

Regular Expression Matching and Operational Semantics

Asiri Rathnayake

Hayo Thielecke

University of Birmingham
Birmingham B15 2TT, United Kingdom

Many programming languages and tools, ranging from `grep` to the Java String library, contain regular expression matchers. Rather than first translating a regular expression into a deterministic finite automaton, such implementations typically match the regular expression on the fly. Thus they can be seen as virtual machines interpreting the regular expression much as if it were a program with some non-deterministic constructs such as the Kleene star. We formalize this implementation technique for regular expression matching using operational semantics. Specifically, we derive a series of abstract machines, moving from the abstract definition of matching to increasingly realistic machines. First a continuation is added to the operational semantics to describe what remains to be matched after the current expression. Next, we represent the expression as a data structure using pointers, which enables redundant searches to be eliminated via testing for pointer equality. From there, we arrive both at Thompson’s lockstep construction and a machine that performs some operations in parallel, suitable for implementation on a large number of cores, such as a GPU. We formalize the parallel machine using process algebra and report some preliminary experiments with an implementation on a graphics processor using CUDA.

1 Introduction

Regular expressions form a minimalistic language of pattern-matching constructs. Originally defined in Kleene’s work on the foundations of computation, they have become ubiquitous in computing. Their practical significance was boosted by Thompson’s efficient construction [13] of a regular expression matcher based on the “lockstep” simulation of a Non-deterministic Finite Automaton (NFA), and the wide use of regular expressions in Unix tools such as `grep` and `awk`.

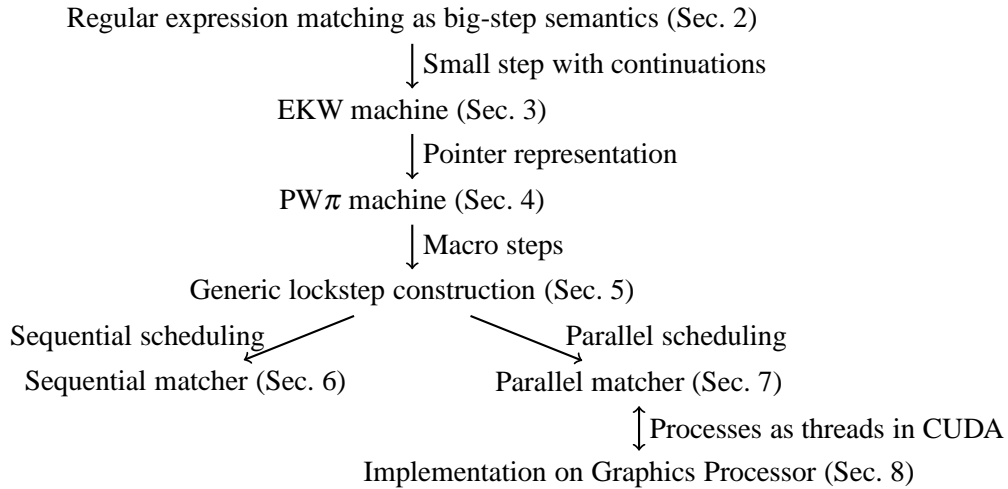
The regular expression matchers used in such tools differ in detail from the implementation of regular expressions used in compiler construction for lexical analysis. In compiling, lexical analyzers are typically built by constructing a Deterministic Finite Automaton (DFA), using one of the standard results of automata theory. The DFA can process input very efficiently, but its construction incurs an additional overhead before any input can be matched. Moreover, the DFA construction only works if the matching language really is a regular language, so that it can be recognized by a DFA. Many matching languages add constructs that take the language beyond what a DFA can recognize, for instance back references. (By abuse of terminology, such extended languages are sometimes still referred to as “regexes”.)

Recently, Cox [5] has given a rational reconstruction of Thompson’s classic NFA matcher in terms of virtual machines. In essence, a regular expression is interpreted on the fly, much as a program in an interpreted programming language. The interpreter is a kind of virtual machine, with a small set of instructions suitable for running regular expressions. For instance, the Kleene star e^* gives a form of non-deterministic loop. Cox emphasizes that the virtual machine approach in the style of Thompson is both flexible and efficient. Once a basic virtual machine for regular expressions is set up, other constructs such as back-references can be added with relative ease. Moreover, the machine is much more efficient than other implementation techniques based on a more naive backtracking interpreter [4], which exhibit

exponential run-time in some cases. Surprisingly, these inefficient matchers are widely used in Java and Perl [4].

In this paper, we formalize the view of regular expression matchers as machines by using tools from programming language theory, specifically operational semantics. We do so starting from the usual definition of regular expressions and their meaning, and then defining increasingly realistic machines.

We first define some preliminaries and recall what it means for a string to match a regular expression in Section 2; from our perspective, matching is a simple form of big-step semantics, and we aim to refine it into a small-step semantics. To do so in Section 3, we introduce a distinction between a current expression and its continuation. We then refine this semantics by representing the regular expression as a syntax tree using pointers in memory (Section 4). Crucially, the pointer representation allows us to compare sub-expressions by pointer equality (rather than structurally). This pointer equality test is needed for the efficient elimination of redundant match attempts, which underlies the general lockstep NFA simulation presented in Section 5. We recover Thompson’s machine as a sequential implementation of the lockstep construction (Section 6). Since the lockstep construction involves simulating many non-deterministic machines in parallel, we then explore a parallel version using some simple process algebra in Section 7. The parallel process semantics is then related to a prototype implementation we have written in CUDA [3] to run on a Graphics Processor Unit (GPU) in Section 8. Section 9 concludes with some future directions. The overall plan of the paper can be visualised as follows:



2 Regular expression matching as a big-step semantics

Let Σ be a finite set, regarded as the input alphabet. We use the following abstract syntax for regular expressions:

$$\begin{aligned}
 e &::= \varepsilon \\
 e &::= a \quad \text{where } a \in \Sigma \\
 e &::= e^* \\
 e &::= e_1 e_2 \\
 e &::= e_1 \mid e_2
 \end{aligned}$$

We let e range over regular expressions, a over characters, and w over strings of characters. The

$$\boxed{e \downarrow w}$$

$$\begin{array}{c}
\frac{e_1 \downarrow w_1 \quad e_2 \downarrow w_2}{(e_1 e_2) \downarrow (w_1 w_2)} \text{ (SEQ)} \quad \frac{}{a \downarrow a} \text{ (MATCH)} \quad \frac{}{\varepsilon \downarrow \varepsilon} \text{ (EPSILON)} \\
\\
\frac{e \downarrow w_1 \quad e^* \downarrow w_2}{e^* \downarrow (w_1 w_2)} \text{ (KLEENE1)} \quad \frac{}{e^* \downarrow \varepsilon} \text{ (KLEENE2)} \\
\\
\frac{e_1 \downarrow w}{(e_1 \mid e_2) \downarrow w} \text{ (ALT1)} \quad \frac{e_2 \downarrow w}{(e_1 \mid e_2) \downarrow w} \text{ (ALT2)}
\end{array}$$

Figure 2.1: Regular expression matching as a big-step semantics

empty string is written as ε . Note that there is also a regular expression constant ε . We also write the sequential composition $e_1 e_2$ as $e_1 \bullet e_2$ when we want to emphasise it as the occurrence of an operator applied to e_1 and e_2 , for instance in a syntax tree. For strings w_1 and w_2 , we write their concatenation as juxtaposition $w_1 w_2$. A single character a is also regarded as a string of length 1.

Our starting point is the usual definition of what it means for a string w to match a regular expression e . We write this relation as $e \downarrow w$, regarding it as a big-step operation semantics for a language with non-deterministic branching $e_1 \mid e_2$ and a non-deterministic loop e^* . The rules are given in Figure 2.1.

Some of our operational semantics will use lists. We write $h :: t$ for constructing a list with head h and tail t . The concatenation of two lists s and t is written as $s @ t$. For example, $1 :: [2] = [1, 2]$ and $[1, 2] @ [3] = [1, 2, 3]$. The empty list is written as $[]$.

3 The EKW machine

The big-step operational semantics of matching in Figure 2.1 gives us little information about how we should attempt to match a given input string w . We define a small-step semantics, called the EKW machine, that makes the matching process more explicit. In the tradition of the SECD machine [7], the machine is named after its components: E for expression, K for continuation, W for word to be matched.

Definition 3.1 A configuration of the EKW machine is of the form $\langle e ; k ; w \rangle$ where e is a regular expression, k is a list of regular expressions, and w is a string. The transitions of the EKW machine are given in Figure 3.1. The accepting configuration is $\langle \varepsilon ; [] ; \varepsilon \rangle$.

Here e is the regular expression the machine is currently focusing on. What remains to the right of the current expression is represented by k , the current continuation. The combination of e and k together is attempting to match w , the current input string.

Note that many of the rules are fairly standard, specifically the pushing and popping of the continuation stack. The machine is non-deterministic. The paired rules with the same current expressions e^* or $(e_1 \mid e_2)$ give rise to branching in order to search for matches, where it is sufficient that one of the branches succeeds.

Theorem 3.2 (Partial correctness) $e \downarrow w$ if and only if there is a run

$$\langle e ; [] ; w \rangle \rightarrow \cdots \rightarrow \langle \varepsilon ; [] ; \varepsilon \rangle$$

$$\boxed{\langle e ; k ; w \rangle \rightarrow \langle e' ; k' ; w' \rangle}$$

$$\langle e_1 \mid e_2 ; k ; w \rangle \rightarrow \langle e_1 ; k ; w \rangle \quad (3.1)$$

$$\langle e_1 \mid e_2 ; k ; w \rangle \rightarrow \langle e_2 ; k ; w \rangle \quad (3.2)$$

$$\langle e_1 e_2 ; k ; w \rangle \rightarrow \langle e_1 ; e_2 :: k ; w \rangle \quad (3.3)$$

$$\langle e^* ; k ; w \rangle \rightarrow \langle e ; e^* :: k ; w \rangle \quad (3.4)$$

$$\langle e^* ; k ; w \rangle \rightarrow \langle \varepsilon ; k ; w \rangle \quad (3.5)$$

$$\langle a ; k ; aw \rangle \rightarrow \langle \varepsilon ; k ; w \rangle \quad (3.6)$$

$$\langle \varepsilon ; e :: k ; w \rangle \rightarrow \langle e ; k ; w \rangle \quad (3.7)$$

Figure 3.1: EKW machine transition steps

Example 3.3 Unfortunately, while Theorem 3.2 ensures that all matching strings are correctly accepted, there is no guarantee that the machine accepts all strings that it should on every run. In fact, there are valid inputs on which the machine may enter an infinite loop; an example is the configuration $\langle a^{**} ; [] ; a \rangle$.

$$\begin{aligned} \langle a^{**} ; [] ; a \rangle &\rightarrow \langle a^* ; [a^{**}] ; a \rangle \\ &\rightarrow \langle \varepsilon ; [a^{**}] ; a \rangle \\ &\rightarrow \langle a^{**} ; [] ; a \rangle \\ &\rightarrow \dots \end{aligned}$$

Such infinite loops can be prevented by backtracking and pruning. However, backtracking implementations can still take a very long time matching expressions like a^{**} to a string consisting of, say, 1000 occurrences of a character a followed by some other b , due to the exponentially increasing search space [4].

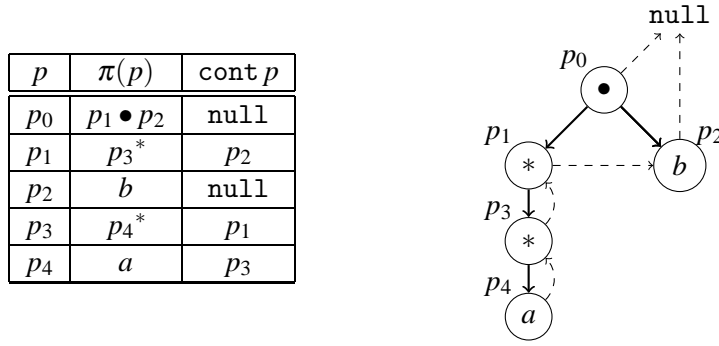
In Thompson's matcher, such loops are avoided by means of redundancy elimination. The matcher checks whether it has encountered the same expression before. Note, however, that "the same" expression is to be taken in the sense of pointer equality rather than structural equality. For instance, the two occurrences of a in $(ab) \mid (ac)$ would be taken as not the same, given their different positions in the syntax tree.

4 The PW π machine

We refine the EKW machine by representing the regular expression as a data structure in a heap π , which serves as the program run by the machine. That way, the machine can distinguish between different positions in the syntax tree.

Definition 4.1 A heap π is a finite partial function from addresses to values. There exists a distinguished address `null`, which is not mapped to any value.

In our setting, the values are syntax tree nodes, represented by an operator from the syntax of regular expressions together with pointers to the tree for the arguments (if any) of the operator. For example, for

Figure 4.1: The regular expression $a^{**} \bullet b$ as a tree with continuation pointers

sequential composition, we have a node containing $(p_1 \bullet p_2)$, where the two pointers p_1 and p_2 point to the trees of the two expressions being composed.

Definition 4.2 We write \otimes for the partial operation of forming the union of two partial functions provided that their domains are disjoint. More formally, let $f_1 : A \rightarrow B$ and $f_2 : A \rightarrow B$ be two partial functions. Then if $\text{dom}(f_1) \cap \text{dom}(f_2) = \emptyset$, the function

$$(f_1 \otimes f_2) : A \rightarrow B$$

is defined as $f_1 \otimes f_2 = f_1 \cup f_2$.

Note that \otimes is the same as the operation $*$ on heaps in separation logic [11], and hence a partial commutative monoid. We avoid the notation $*$ as it could be confused with the Kleene star. As in separation logic, we use \otimes to describe data structures with pointers in memory.

Definition 4.3 We write $\pi, p \models e$ if p points to the root node of a regular expression e in a heap π . The relation is defined by induction on e as follows:

$$\begin{array}{ll}
 \pi, p \models a & \text{if } \pi(p) = a \\
 \pi, p \models \varepsilon & \text{if } \pi(p) = \varepsilon \\
 \pi, p \models (e_1 \mid e_2) & \text{if } \pi = \pi_0 \otimes \pi_1 \otimes \pi_2 \wedge \pi_0(p) = (p_1 \mid p_2) \\
 & \wedge \pi_1, p_1 \models e_1 \wedge \pi_2, p_2 \models e_2 \\
 \pi, p \models (e_1 e_2) & \text{if } \pi = \pi_0 \otimes \pi_1 \otimes \pi_2 \wedge \pi_0(p) = (p_1 \bullet p_2) \\
 & \wedge \pi_1, p_1 \models e_1 \wedge \pi_2, p_2 \models e_2 \\
 \pi, p \models e_1^* & \text{if } \pi = \pi_0 \otimes \pi_1 \wedge \pi_0(p) = p_1^* \wedge \pi_1, p_1 \models e_1
 \end{array}$$

Here the definition of $\pi, p \models e$ precludes any cycles in the child pointer chain.

As an example, consider the regular expression $e = a^{**}b$. A π and p_0 such that $\pi, p_0 \models e$ is given by the table in Figure 4.1. The tree structure, represented by the solid arrows, is drawn on the right.

$$\boxed{p \longrightarrow q \text{ or } p \xrightarrow{a} q \text{ relative to } \pi}$$

$$\begin{aligned}
p &\longrightarrow p_1 && \text{if } \pi(p) = p_1 \mid p_2 \\
p &\longrightarrow p_2 && \text{if } \pi(p) = p_1 \mid p_2 \\
p &\longrightarrow p_1 && \text{if } \pi(p) = p_1 \bullet p_2 \\
p &\longrightarrow p_1 && \text{if } \pi(p) = p_1^* \\
p &\longrightarrow p_2 && \text{if } \pi(p) = p_1^* \text{ and } \text{cont } p = p_2 \\
p &\longrightarrow p_1 && \text{if } \pi(p) = \varepsilon \text{ and } \text{cont } p = p_1 \\
p &\xrightarrow{a} p' && \text{if } \pi(p) = a \text{ and } p' = \text{cont } p
\end{aligned}$$

Figure 4.2: PW π transitions

Definition 4.4 Let cont be a function

$$\text{cont} : \text{dom}(\pi) \rightarrow (\text{dom}(\pi) \cup \{\text{null}\})$$

We write $\pi \models \text{cont}$ if

- If $\pi(p) = (p_1 \mid p_2)$, then $\text{cont } p_1 = \text{cont } p$ and $\text{cont } p_2 = \text{cont } p$
- If $\pi(p) = (p_1 \bullet p_2)$, then $\text{cont } p_1 = p_2$ and $\text{cont } p_2 = \text{cont } p$
- If $\pi(p) = (p_1)^*$, then $\text{cont } p_1 = p$
- $\text{cont } p_0 = \text{null}$, where p_0 is the pointer to the root of the syntax tree.

The function cont is uniquely determined by the tree structure layed out in π , and it is easy to compute by a recursive tree walk. We elide it when it is clear from the context, assuming that π always comes equipped with a cont such that $\pi \models \text{cont}$. By treating cont as a function, we have not committed to a particular implementation; for instance cont could be represented as a hash table indexed by pointer values, or it could be added as another pointer field to the nodes in the heap.

In the graphical representation in Figure 4.1, dashed arrows represent cont . In particular, note the cycle leading downward from p_1 and up again via dashed arrows. Following such a cycle could lead to infinite loops as for the EKW machine in Example 3.3.

Definition 4.5 The PW π machine is defined as follows. Transitions of this machine are always relative to some heap π , which does not change during evaluation. We elide π if it is clear from the context. Configurations of the machine are of the form $\langle p ; w \rangle$, where p is a pointer in π and w is a string of input symbols. Given the transition relation between pointers defined in Figure 4.2, the machine has the following transitions:

$$\frac{p \xrightarrow{a} q}{\langle p ; a w \rangle \rightarrow \langle q ; w \rangle} \qquad \frac{p \longrightarrow q}{\langle p ; w \rangle \rightarrow \langle q ; w \rangle}$$

The accepting state of the machine is $\langle \text{null} ; \varepsilon \rangle$. That is, both the continuation and the remaining input have been consumed.

Example 4.6 For a regular expression $e = a^{**}b$, let π and p_0 be such that $\pi, p_0 \models e$. See Figure 4.1 for the representation of π as a tree with pointers. The diagram below illustrates two possible executions of the $PW\pi$ machine against inputs e and aab .

Execution - 1: Infinite loop

$$\begin{aligned} & \langle p_0 ; aab \rangle \\ \longrightarrow & \langle p_1 ; aab \rangle \\ \longrightarrow & \langle p_3 ; aab \rangle \\ \longrightarrow & \langle p_1 ; aab \rangle \\ \longrightarrow & \langle p_3 ; aab \rangle \\ \longrightarrow & \langle p_1 ; aab \rangle \\ \longrightarrow & \langle p_3 ; aab \rangle \\ \longrightarrow & \langle p_1 ; aab \rangle \\ \longrightarrow & \langle p_3 ; aab \rangle \\ \longrightarrow & \langle p_1 ; aab \rangle \\ \longrightarrow & \dots \end{aligned}$$

Execution - 2: Successful match

$$\begin{aligned} & \langle p_0 ; aab \rangle \\ \longrightarrow & \langle p_1 ; aab \rangle \\ \longrightarrow & \langle p_3 ; aab \rangle \\ \longrightarrow & \langle p_4 ; aab \rangle \\ \longrightarrow & \langle p_3 ; ab \rangle \\ \longrightarrow & \langle p_4 ; ab \rangle \\ \longrightarrow & \langle p_3 ; b \rangle \\ \longrightarrow & \langle p_1 ; b \rangle \\ \longrightarrow & \langle p_2 ; b \rangle \\ \longrightarrow & \langle \text{null} ; \varepsilon \rangle \end{aligned}$$

Theorem 4.7 (Simulation) Let π be a heap such that $\pi, p \models e$. Then there is a run of the EKW machine of the form

$$\langle e ; [] ; w \rangle \rightarrow \dots \rightarrow \langle \varepsilon ; [] ; \varepsilon \rangle$$

if and only if there is a run of the $PW\pi$ machine of the form

$$\langle p ; w \rangle \rightarrow \dots \rightarrow \langle \text{null} ; \varepsilon \rangle$$

One needs to show that each step of the EKW machine can be simulated by the $PW\pi$ machine and vice versa. The invariant in this simulation is that the stack k in the EKW machine can be reconstructed by following the chain of pointers in the heap of the $PW\pi$ machine via the following function:

$$\begin{aligned} \text{stack } p &= [] && \text{if } \text{cont } p = \text{null} \\ \text{stack } p &= e :: (\text{stack } q) && \text{if } q = \text{cont } p \neq \text{null} \\ &&& \text{and } \pi, q \models e \end{aligned}$$

5 The lockstep construction in general

As we have seen, the $PW\pi$ machine is built from two kinds of steps. Pointers can be evolved via $p \longrightarrow q$ by moving in the syntax tree without reading any input. When a node for a constant is reached, it can be matched to the first character in the input via a step $p \xrightarrow{a} q$.

Definition 5.1 Let $S \subseteq \text{dom}(\pi) \cup \{\text{null}\}$ be a set of pointers. We define the evolution $\Box S$ of S as the following set:

$$\Box S = \{q \in \text{dom}(\pi) \mid \exists p \in S. p \longrightarrow^* q \wedge \exists a. \pi(q) = a\}$$

Forming $\Box S$ is similar to computing the ε -closure in automata theory. However, this operation is not a closure operator, because $S \subseteq \Box S$ does not hold in general. When one computes $\Box S$ incrementally, elements are removed as well as added. Avoiding infinite loops by adding and removing the same element is the main difficulty in the computation.

We define a transition relation analogous to Definition 4.5, but as a deterministic relation on *sets* of pointers. We refer to these as macro steps, as they assume the computation of $\Box S$ as given in a single step, whereas an implementation needs to compute it incrementally.

Definition 5.2 (Lockstep transitions) Let $S, S' \subseteq \text{dom}(\pi) \cup \{\text{null}\}$ be sets of pointers.

$$\begin{aligned} S &\Longrightarrow S' && \text{if } S' = \Box S \\ S &\xRightarrow{a} S' && \text{if } S' = \{q \in \text{dom}(\pi) \mid \exists p \in S. p \xrightarrow{a} q\} \end{aligned}$$

A set of pointers is first evolved from S to $\Box S$. Then, moving from a set of pointers $\Box S$ to S' via $\Box S \xRightarrow{a} S'$ advances the state of the machine by advancing all pointers that can match a to their continuations. All other pointers are deleted as unsuccessful matches.

Definition 5.3 (Generic lockstep machine) The generic lockstep machine has configurations of the form $\langle S; w \rangle$. Transitions are defined using Definition 5.2:

$$\frac{S \xRightarrow{a} S'}{\langle S; a w \rangle \Rightarrow \langle S'; w \rangle} \quad \frac{S \Longrightarrow S'}{\langle S; w \rangle \Rightarrow \langle S'; w \rangle}$$

Accepting states of the machine are of the form $\langle S; \varepsilon \rangle$, where $\text{null} \in S$.

Theorem 5.4 For a heap $\pi, p \models e$ there is a run of the $\text{PW}\pi$ machine:

$$\langle p; w \rangle \rightarrow \cdots \rightarrow \langle \text{null}; \varepsilon \rangle$$

if and only if there is a run of the lockstep machine

$$\langle \{p\}; w \rangle \Rightarrow \cdots \Rightarrow \langle S; \varepsilon \rangle$$

for some set of pointers S with $\text{null} \in S$.

6 The sequential lockstep machine

The sequential lockstep machine maintains two lists of pointers c, n corresponding to pointers being incrementally evolved within the current macro step and pointers to be evolved in the next macro step. Another pointer list t is maintained which provides support for redundancy elimination, we also introduce an auxiliary function $\psi(p, l_1, l_2)$ to aid in this regard:

Definition 6.1 The auxiliary function $\psi(p, l_1, l_2)$ is defined as:

$$\begin{aligned} \psi(p, l_1, l_2) &= p :: l_1 \text{ if } p \notin l_1 @ l_2 \\ \psi(p, l_1, l_2) &= l_1 \text{ if } p \in l_1 @ l_2 \end{aligned}$$

$\langle c; t; n; w \rangle \rightarrow \langle c'; t'; n'; w' \rangle$		
$\langle p :: c; t; n; w \rangle \rightarrow \langle c'; p :: t; n; w \rangle$	if $\pi(p) = p' \mid p''$ where $c' = \psi(p'', \psi(p', c, t), t)$	
$\langle p :: c; t; n; w \rangle \rightarrow \langle c'; p :: t; n; w \rangle$	if $\pi(p) = p' \bullet p''$ where $c' = \psi(p', c, t)$	
$\langle p :: c; t; n; w \rangle \rightarrow \langle c'; p :: t; n; w \rangle$	if $\pi(p) = (p')^*$ where $c' = \psi(\text{cont } p, \psi(p, c, t), t)$	
$\langle p :: c; t; n; w \rangle \rightarrow \langle c'; p :: t; n; w \rangle$	if $\pi(p) = \varepsilon$ where $c' = \psi(\text{cont } p, c, t)$	
$\langle p :: c; t; n; aw \rangle \rightarrow \langle c; t; n; aw \rangle$	if $p = \text{null}$	
$\langle p :: c; t; n; aw \rangle \rightarrow \langle c; p :: t; n'; aw \rangle$	if $\pi(p) = a$ where $n' = \psi(\text{cont } p, n, [])$	
$\langle p :: c; t; n; aw \rangle \rightarrow \langle c; p :: t; n; aw \rangle$	if $\pi(p) = b$	
$\langle []; t; n; aw \rangle \rightarrow \langle n; []; []; w \rangle$	if $n \neq []$	
$\langle p :: c; t; n; \varepsilon \rangle \rightarrow \langle c; p :: t; n; \varepsilon \rangle$	if $\pi(p) = a$	

Figure 6.1: Sequential lockstep machine with redundancy elimination

Definition 6.2 The redundancy-eliminating sequential lockstep machine has configurations of the form $\langle c; t; n; w \rangle$. Its transitions are given in figure 6.1. The accepting states are of the form $\langle \text{null} :: c'; t'; n'; \varepsilon \rangle$

We regard this machine as a rational reconstruction of Thompson’s matcher [13] in the light of Cox’s elucidation as a virtual machine [5]. This machine uses a sequential schedule for incrementally evolving pointers, keeping a list of pointers that have been evolved already to prevent loops and search space explosion. However, our main interest is in performing this computation in parallel.

7 Parallel lockstep semantics

We now define an operational semantics where each pointer is given a dedicated thread for evolving it. Our motivation is to leverage the large number of cores and hence threads available on GPUs. The semantics in this section is intended as an idealization of the implementation described in Section 8 below, capturing the essentials of the computation while abstracting from implementation details.

To describe the parallel computation, we define a simple process calculus. Its transition rules are given in Figure 7.1. Most of our calculus is a subset of CCS [8], with one-to-one directional message passing and parallel composition. However, we also need an n -way synchronization with a synchronous transition inspired by Synchronous CCS [9].

We let M range over processes, p over pointers that may be sent as asynchronous messages, and a

$$\boxed{M \longrightarrow M'} \quad \frac{M_1 \longrightarrow M_2}{M_1 \parallel M_3 \longrightarrow M_2 \parallel M_3} \text{ (PAR)} \quad \frac{}{((p.M) \parallel \bar{p}) \longrightarrow M} \text{ (SEND)}$$

$$\boxed{M \xrightarrow{a} M'} \quad \frac{M' \not\equiv (\$a.M') \parallel M''' \quad M' \not\rightarrow}{(\$a.M_1 \parallel \dots \parallel \$a.M_n \parallel M') \xrightarrow{a} (M_1 \parallel \dots \parallel M_n)} \text{ (SYNC)}$$

Figure 7.1: Process calculus

over input symbols, which may be used for n -way synchronisation. The syntax of processes is as follows:

$$\begin{aligned}
M ::= & \bar{p} \\
& | M \parallel M \\
& | p.M \\
& | \$a.M
\end{aligned}$$

We impose some structural congruences \equiv , identifying terms up to associativity and commutativity of parallel composition \parallel . Process transitions can be interleaved with rule PAR.

We have CCS-style handshake communication in rule SEND. Here $p.M$ receives the message \bar{p} and proceeds with M afterwards. Note that receivers of the form $p.M$ are not replicated (in the pi-calculus sense [10]), so that each communication consumes the receiver. This behaviour is essential, as the processes we generate could become trapped in an infinite loop otherwise.

We also have an n -way synchronisation SYNC. This rule is the most complex, and it is needed to implement matching to input once all pointers have been evolved. The idea is as follows:

- The current process is factorized into those processes that are of the form $\$a.M_j$ and an M' comprising everything else.
- There are no further \longrightarrow transitions inside M' , written as $M' \not\rightarrow$.
- If these conditions are met, then all the processes waiting to participate in an n -way synchronisation on a are advanced in one synchronous step.
- The remaining processes in M' are discarded in the same step.

Rules in this style, in which a number of processes are advanced in a single step, are sometimes referred to as “lockstep” [9]. Indeed, we use this rule to implement the lockstep matching of regular expressions in the sense of Thompson and Cox. (In practice, this rule may require a little ad-hoc protocol to implement on a given architecture.)

We translate each expression pointer p in the heap π into a process $\llbracket p \rrbracket \pi$ as follows:

$$\begin{aligned}
\llbracket p \rrbracket \pi &= p.(\bar{q_1} \parallel \bar{q_2}) & \text{if } \pi(p) &= (q_1 \mid q_2) \\
\llbracket p \rrbracket \pi &= p.\bar{q_1} & \text{if } \pi(p) &= (q_1 \bullet q_2) \\
\llbracket p \rrbracket \pi &= p.(\bar{q_1} \parallel \bar{q_2}) & \text{if } \pi(p) &= q_1^* \text{ and } \text{cont } p = q_2 \\
\llbracket p \rrbracket \pi &= p.\bar{q} & \text{if } \pi(p) &= \epsilon \text{ and } \text{cont } p = q \\
\llbracket p \rrbracket \pi &= p.\$a.\bar{q} & \text{if } \pi(p) &= a \text{ and } \text{cont } p = q
\end{aligned}$$

Intuitively, for each internal node in the expression tree identified by the pointer p , we create a dedicated little process that listens on a channel uniquely corresponding to p . For simplicity, we use the same name for the channel as for the pointer. The process may be activated by messages \bar{p} sent to it, and it may send such messages itself. These messages trigger a chain reaction that evolve the current pointer set of a macro step. There is no need for these messages to be externally visible, as their only purpose is to wake up their unique recipient. A process $p.M$ listening for \bar{p} is consumed by the transition that receives the message. Processes for nodes that point to input characters a at the leaves of the expression tree use a different form of communication. All these nodes synchronize on the input symbol. The symbol a is visible in the resulting synchronous transition step \xrightarrow{a} , because we need it to agree with the next input symbol.

If $\text{dom}(\pi) = \{p_1, \dots, p_n\}$, we define the translation $\llbracket \pi \rrbracket$ as the translation of all its pointers:

$$\llbracket p_1 \rrbracket \pi \parallel \dots \parallel \llbracket p_n \rrbracket \pi$$

If the input string is not empty, let a be the first character, so that $aw' = w$. The parallel machine launches processes for all the nodes in the tree, and sends a message to the process for the root. The resulting process makes a number of asynchronous transitions, followed by a synchronous move for a :

$$\llbracket \pi \rrbracket \parallel \bar{p} \longrightarrow \dots \longrightarrow \xrightarrow{a} M$$

All these steps together represent one macro step. The machine then repeats the above with the next symbol a' and M

$$\llbracket \pi \rrbracket \parallel M \longrightarrow \dots \longrightarrow \xrightarrow{a'} M'$$

The machine accepts if the remaining input is empty and the current process is of the form

$$\overline{\text{null}} \parallel M$$

Example 7.1 For $e = a^{**}b$, let π and p_0 be such that $\pi, p_0 \models e$. See Figure 4.1 for the representation of π as a tree with pointers. Translating the tree structure to parallel processes gives us:

$$\llbracket \pi \rrbracket = (p_0 \cdot \bar{p}_1) \parallel p_1 \cdot (\bar{p}_3 \parallel \bar{p}_2) \parallel p_2 \cdot \$b \cdot \overline{\text{null}} \parallel p_3 \cdot (\bar{p}_4 \parallel \bar{p}_1) \parallel p_4 \cdot \$a \cdot \bar{p}_3$$

Assume an input string of aab . We have the pointer evolution as follows:

$$\begin{aligned} & \bar{p}_0 \parallel \llbracket \pi \rrbracket \\ \longrightarrow & \bar{p}_0 \parallel p_0 \cdot \bar{p}_1 \parallel p_1 \cdot (\bar{p}_3 \parallel \bar{p}_2) \parallel p_2 \cdot \$b \cdot \overline{\text{null}} \parallel p_3 \cdot (\bar{p}_4 \parallel \bar{p}_1) \parallel p_4 \cdot \$a \cdot \bar{p}_3 \\ \longrightarrow & \bar{p}_1 \parallel p_1 \cdot (\bar{p}_3 \parallel \bar{p}_2) \parallel p_2 \cdot \$b \cdot \overline{\text{null}} \parallel p_3 \cdot (\bar{p}_4 \parallel \bar{p}_1) \parallel p_4 \cdot \$a \cdot \bar{p}_3 \\ \longrightarrow & \bar{p}_3 \parallel \bar{p}_2 \parallel p_2 \cdot \$b \cdot \overline{\text{null}} \parallel p_3 \cdot (\bar{p}_4 \parallel \bar{p}_1) \parallel p_4 \cdot \$a \cdot \bar{p}_3 \\ \longrightarrow & \bar{p}_3 \parallel \$b \cdot \overline{\text{null}} \parallel p_3 \cdot (\bar{p}_4 \parallel \bar{p}_1) \parallel p_4 \cdot \$a \cdot \bar{p}_3 \\ \longrightarrow & \$b \cdot \overline{\text{null}} \parallel \bar{p}_4 \parallel \bar{p}_1 \parallel p_4 \cdot \$a \cdot \bar{p}_3 \\ \longrightarrow & \$b \cdot \overline{\text{null}} \parallel \bar{p}_1 \parallel \$a \cdot \bar{p}_3 \end{aligned}$$

Since no more micro transitions are possible, we have reached the n -way synchronization point:

$$\$b \cdot \overline{\text{null}} \parallel \bar{p}_1 \parallel \$a \cdot \bar{p}_3 \xrightarrow{a} \bar{p}_3$$

Now we feed the residual messages back into a fresh $\llbracket \pi \rrbracket$:

$$\begin{aligned}
& \overline{p_3} \parallel \llbracket \pi \rrbracket \\
& \longrightarrow \overline{p_3} \parallel p_0 \cdot \overline{p_1} \parallel p_1 \cdot (\overline{p_3} \parallel \overline{p_2}) \parallel p_2 \cdot \$b \cdot \overline{\text{null}} \parallel p_3 \cdot (\overline{p_4} \parallel \overline{p_1}) \parallel p_4 \cdot \$a \cdot \overline{p_3} \\
& \longrightarrow p_0 \cdot \overline{p_1} \parallel p_1 \cdot (\overline{p_3} \parallel \overline{p_2}) \parallel p_2 \cdot \$b \cdot \overline{\text{null}} \parallel \overline{p_4} \parallel \overline{p_1} \parallel p_4 \cdot \$a \cdot \overline{p_3} \\
& \longrightarrow p_0 \cdot \overline{p_1} \parallel p_1 \cdot (\overline{p_3} \parallel \overline{p_2}) \parallel p_2 \cdot \$b \cdot \overline{\text{null}} \parallel \overline{p_1} \parallel \$a \cdot \overline{p_3} \\
& \longrightarrow p_0 \cdot \overline{p_1} \parallel \overline{p_3} \parallel \overline{p_2} \parallel p_2 \cdot \$b \cdot \overline{\text{null}} \parallel \$a \cdot \overline{p_3} \\
& \longrightarrow p_0 \cdot \overline{p_1} \parallel \overline{p_3} \parallel \$b \cdot \overline{\text{null}} \parallel \$a \cdot \overline{p_3} \\
& \xrightarrow{a} \overline{p_3} \\
& \longrightarrow \dots \\
& \xrightarrow{b} \overline{\text{null}}
\end{aligned}$$

Therefore, we have received a $\overline{\text{null}}$ while the input string has become empty, resulting in a successful match.

We need to prove that the construction above can correctly evolve and match any set of pointers. Let $S = \{p_1, \dots, p_n\} \subseteq \text{dom}(\pi) \cup \{\text{null}\}$ be a set of pointers in the heap. We define

$$\overline{S} = \overline{p_1} \parallel \dots \parallel \overline{p_n}$$

to represent this set as a parallel composition of messages.

Theorem 7.2 Let $S, S' \subseteq \text{dom}(\pi) \cup \{\text{null}\}$. We have

$$S \Longrightarrow \xRightarrow{a} S'$$

if and only if

$$\overline{S} \parallel \llbracket \pi \rrbracket \longrightarrow^* \xrightarrow{a} \overline{S'}$$

Moreover, each \longrightarrow transition sequence starting from $\overline{S} \parallel \llbracket \pi \rrbracket$ is finite.

Theorem 7.2 assures us that the parallel operational semantics correctly implements the lockstep construction. The pointers p in the tree, represented as processes \overline{p} , are evolved in parallel. Although this evolution is non-deterministic, its end result is determinate. Moreover, the cycles in the pointer chain do not lead to cyclic processes looping forever, since each receiving process becomes inactive once the node has been visited.

The correctness proof of the parallel implementation relies on a factorisation of the processes into four components. At each step i , we have:

- A set S_i of pointers, indicating nodes that should be evolved.
- A heap of receivers $\pi_i \subseteq \pi$, representing nodes that have not been visited in the current macro step.
- A set E_i of evolved nodes, whose process representations are of the form ready to match a character.
- A parallel composition D_i of messages to nodes that have already been processed.

Let E be a set of pointers $E = \{p_1, \dots, p_n\}$ such that $\pi(p_j) = a_j$ and $\text{cont } p_j = q_j$. We write

$$\$E = \$a_1.q_1 \parallel \dots \parallel \$a_n.q_n$$

We need to consider transition sequences of the form

$$\begin{array}{c} \overline{S_0} \parallel \llbracket \pi_0 \rrbracket \parallel \$E_0 \parallel D_0 \\ \longrightarrow \\ \vdots \\ \longrightarrow \overline{S_n} \parallel \llbracket \pi_n \rrbracket \parallel \$E_n \parallel D_n \end{array}$$

where $\pi_0 = \pi$ and $E_0 = \emptyset$. The invariant we need to establish for all transition steps consists of:

$$\begin{aligned} \Box S_0 &= \Box (S_i \cap \text{dom}(\pi_i)) \cup E_i \\ \Box R_i &\subseteq \Box (S_i \cap \text{dom}(\pi_i)) \cup E_i \\ \{p \mid \exists D. D_i \equiv (\overline{p} \parallel D)\} &\subseteq S_i \cup R_i \end{aligned}$$

where $R_i = \text{dom}(\pi) \setminus \text{dom}(\pi_i)$. The factorization of processes at each step and the invariant are verified by case analysis on the kind of node $\pi(p)$ and hence the possible \longrightarrow steps that its translation $\llbracket p \rrbracket \pi$ can make using the rules from Figure 7.1.

In the final configuration we have $S_n \cap \text{dom}(\pi_n) = \emptyset$. Hence,

$$\begin{aligned} \Box S_0 &= \Box (S_n \cap \text{dom}(\pi_n)) \cup E_n \\ &= \Box \emptyset \cup E_n \\ &= E_n \end{aligned}$$

Therefore, we have $\Box S_0 = E_n$, as required. From that configuration, there can only be an \xrightarrow{a} transition, exactly matching the generic lockstep transition $S \Longrightarrow \xrightarrow{a} S'$.

8 Implementation on a GPU

As a proof of concept, we have written a simple regular expression matcher where the evolution of pointers is performed in parallel on a GPU.¹ Programming the GPU was done via CUDA [3]. The main points are:

- The regular expression is parsed, and the syntax tree nodes are packed into an array d . This array represents our heap π . A second pass through the syntax tree performs the wiring of continuation pointers, corresponding to cont .
- Two integer vectors c, n of the same size as the regular expression vector above are created. Here a value of t - the macro step count, on $c[i]$ implies that regular expression $d[i]$ is to be simulated within the current macro step. On the other hand a value of $-t$ on $c[i]$ implies that the corresponding regular expression has already been simulated for the current macro step. This protocol realizes the semantics of a process being consumed once it has received a message. The vector n is used to collect those search attempts which are able to match the current input character. A value of $t+1$ on $n[j]$ indicates that the regular expression $d[j]$ is to be simulated on the next macro step.

¹The code is available at <http://www.cs.bham.ac.uk/~hxt/research/regexp.shtml>.

- Each regular expression node $d[i]$ is assigned a GPU thread. This GPU thread is responsible for conditionally simulating the regular expression $d[i]$ at each invocation (depending on $c[i]$ value). While simulating an expression, a GPU thread might schedule another GPU thread / expression $d[j]$ by setting $c[j]$ to t (this could happen for an example in the case of $e = e_1 \bullet e_2$). Note that one thread scheduling another thread via the c vector corresponds to the sending of a message \bar{p} from one process to another.
- At each invocation of the GPU threads (called a *kernel launch* in CUDA terminology), each thread which performs a successful simulation updates either of two shared flags which indicate if there were more threads activated on the c or n vectors during the current invocation. A macro transition involves swapping the c and n vectors while incrementing the t counter. It corresponds to the n -way synchronization transition.
- The initial state of the machine has only $d[0]$, the root node process, scheduled for simulation.

However, note that this description corresponds to a minimalistic GPU-based parallel lockstep machine and does not yet incorporate any optimizations from the literature [14], such as *persistent threads* and *tasks queues*.

9 Conclusions

We have derived regular expression matchers as abstract machines. In doing so, we have used a number of concepts and techniques from programming language theory. The EKW machine zooms in on a current expression while maintaining a continuation for keeping track of what to do next. In that sense, the machine is a distant relative of machines for interpreting lambda terms, such as the SECD machine [7] or the CEK machine [6]. On the other hand, regular expressions are a much simpler language to interpret than lambda calculus, so that continuations can be represented by a single pointer into the tree structure (or to machine code in Thompson's original implementation). While the idea of continuations as code pointers is sometimes advanced as a helpful intuition, the representation of continuations in CPS compiling [1] is more complex, involving an environment pointer as well. To represent pointers and the structures they build up, we found it convenient to use a small fragment of separation logic [11], given by just the separating conjunction and the points-to-predicate. (They are written as \otimes and $\pi(p) = e$ above, to avoid clashes with other notation.) A similar use of these connectives to describe trees in the setting of abstract machines was used in our earlier work on B+trees [12]. Here we translate a tree-shaped data structure into a network of processes that communicate in a cascade of messages mirroring the pointers in the tree structure. The semantics of the processes is inspired by the process algebra literature [8, 9, 10]. One reason why a process algebra is suitable for formalizing the lockstep construction with redundancy elimination is that receiving processes are eliminated once they have received a message; they are used linearly, and so are reminiscent of linearly-used continuations [2].

We intend to extend both the process algebra view and our CUDA implementation, while maintaining a close correspondence between them. Regular expression matching is an instance of irregular parallel [14] processing on a GPU, which presents some optimization problems. At the moment, the parallel processing power of the GPU cores is not exercised, as each thread does little more than access the expression tree and activate threads for other nodes. We expect the load on the GPU cores to become more significant when more expensive constructs such as back-references (known to be NP-hard) are added to our matching language. It remains to be seen whether a GPU implementation will become more efficient than a sequential CPU-based one, particularly as the number of GPU cores continues to

increase (it is currently in the hundreds of cores). More generally, the operational semantics and abstract machine approach may be fruitful for reasoning about other forms of General Purpose Graphics Processing Unit (GPGPU) programming.

References

- [1] Andrew Appel (1992): *Compiling with Continuations*. Cambridge University Press.
- [2] Josh Berdine, Peter W. O'Hearn, Uday Reddy & Hayo Thielecke (2002): *Linear Continuation Passing. Higher-order and Symbolic Computation* 15(2/3), pp. 181–208, doi:10.1023/A:1020891112409.
- [3] NVIDIA Corporation (2011): *What is CUDA?* Available at http://www.nvidia.com/object/what_is_cuda_new.html.
- [4] Russ Cox (2007): *Regular Expression Matching Can Be Simple And Fast (but is slow in Java, Perl, PHP, Python, Ruby, ...)*. Available at <http://swtch.com/~rsc/regexp/regexp1.html>.
- [5] Russ Cox (2009): *Regular Expression Matching: the Virtual Machine Approach*. Available at <http://swtch.com/~rsc/regexp/regexp2.html>.
- [6] Matthias Felleisen & Daniel P. Friedman (1986): *Control operators, the SECD-machine, and the λ -calculus*. In M. Wirsing, editor: *Formal Description of Programming Concepts*, North-Holland, pp. 193–217.
- [7] Peter J. Landin (1964): *The Mechanical Evaluation of Expressions*. *The Computer Journal* 6(4), pp. 308–320.
- [8] Robin Milner (1980): *A Calculus of Communicating Systems*. *Lecture Notes in Computer Science* 92, doi:10.1007/3-540-10235-3, Springer.
- [9] Robin Milner (1983): *Calculi for Synchrony and Asynchrony*. *Theoretical Computer Science* 25, pp. 267–310.
- [10] Robin Milner (1999): *Communicating and Mobile Systems: The Pi Calculus*. Cambridge University Press.
- [11] John C. Reynolds (2002): *Separation Logic: A Logic for Shared Mutable Data Structures*. In: *Logic in Computer Science (LICS)*, IEEE, pp. 55–74, doi:10.1109/LICS.2002.1029817.
- [12] Alan P. Sexton & Hayo Thielecke (2008): *Reasoning about B+ Trees with Operational Semantics and Separation Logic*. In: *Twenty-fourth Conference on the Mathematical Foundations of Programming Semantics (MFPS24)*, Electronic Notes in Theoretical Computer Science, pp. 355–369, doi:10.1016/j.entcs.2008.10.021.
- [13] Ken Thompson (1968): *Programming Techniques: Regular expression search algorithm*. *Communications of the ACM* 11(6), pp. 419–422, doi:10.1145/363347.363387.
- [14] Stanley Tzeng, Anjul Patney & John D. Owens (2010): *Task Management for Irregular-Parallel Workloads on the GPU*. In: *High Performance Graphics*, Eurographics Association, pp. 29–37.